Trade-Off between Factors Influencing Quality of the Summary

Martin Soubbotin, Sergei Soubbotin
FreeText Software Technologies, Inc.
http://www.powertextsolutions.com
{martin, sergei}@ powertextsolutions.com

Abstract

Our summarization approach is based on the assumption that quality of the summary is influenced by a set of factors, dependent on lexical and grammatical features of text units selected and arranged while composing the summary. The system has been developed with taking into account six factors influencing the final quality: compliance with the genre "summary", relevance, focusing, compliance with the requested granularity, topic coverage, and cohesion-based ordering. Preliminary selected text units get positive or negative scores depending on the relative impact of features that are responsible for the respective factors. Balancing the impacts of different text features was performed on empirical basis. After initial ranking of text units by the first four factors, they are ordered so that to achieve higher coverage and cohesion. The top-ranked passages, whose total words number does not exceed the required size, are submitted as the resulting summary.

Introduction

Our decision to participate in the 2005 Document Understanding Conference was motivated by opportunity to test the capabilities of our approach to unstructured text processing aimed at deriving and organizing information on a topic of interest.

We have applied this approach in several multi-document summarization systems, producing various kinds of free-style and structured reports on a topic by automatic compiling information from large documents collections. [1]

We consider our multi-document summarization technology as part of a larger set of techniques aimed at most complete ("exhaustive") mining of information from unstructured text documents. In the framework of this technological approach, extracted text items of various kinds are used in subsequent procedures intended to arrange, as well as to transform and to infer new items from them. [2]

The DUC 2005 summarization task was of special interest for us in view of its elaborated set of quality requirements. This prompted us to develop a new variant of our summarization approach that is methodologically based on establishing relationship between the features of extracted text units and the quality of the resulting summary.

In this paper, we first discuss the type of text units from which a summary can be composed (Section 1). In Section 2, we introduce and discuss the notion of factors influencing the quality of the resulting summary. In Sections 3-9, we consider specific quality factors and the respective lexical and grammatical features of documents passages. In the concluding Sections 10-12, we discuss results obtained by our system, possible implications to summaries evaluation, as well as our future work.

1. The Extracted Units

The multi-document summary is composed of text units that are isolated from original documents and fall within entirely different surrounding, thus forming a new text.

Hence, the important precondition of our approach is self-sufficiency of extracted text units, their understandability and readability beyond their original context. No dangling anaphora referring to their original surrounding can be tolerated.

We use certain rules for identifying pronominal anaphora resolved in the same sentence. But if such anaphora refer to the preceding sentence in the source document, then a larger text unit, containing the referred sentence, is considered for inclusion into the summary.

Similarly, we attach the preceding sentence in case of anaphora represented by nouns (while developing our summarization systems we tested certain rules for recognizing such situations).

Thus, the extracted text units consist of one or more sentences. We will use the term "passage" when referring to the extracted units.

Lexical expressions that served for organizing original texts ("Besides that", "Yet", "Nonetheless", "Moreover", "Indeed", "In other words", "In addition", etc.) must be cleaned from the extracted passages. Removal of such expressions (usually at the beginning of the passage) does not affect the correct grammatical structure of the sentence.

Further, our system removes any mentioning of time and date that cannot be interpreted precisely by a reader (“this Monday”, “last month”, “yesterday”, etc.).

The criteria for selecting a passage as a candidate for including into the summary are some minimal indications of its relevance (they will be considered in Section 4).

2. Factors Influencing the Quality

We assume that the quality of a summary is influenced by a set of factors dependent on lexical and grammatical features of text units that are selected and arranged while composing the summary.

We consider six factors influencing the final quality: compliance with the genre "summary", relevance, focusing, compliance with the requested granularity, topic coverage and cohesion-based ordering.

Lexical features responsible for the respective factors include presence or absence of items from the title/narrative text, their derivatives, co-occurring words, as well as presence of words/phrases from multiple embedded vocabularies. Grammatical features are such as the use of nouns in the plural form and presence of some kinds of pronouns.
Some features can be identified in a passage regardless of other passages. There are also features that can be recognized only in correlation to features of other passages, such as repeating term or a newly occurring term from a certain category (e.g., new country name).
To every factor-related feature, positive or negative weights are assigned. The core of this methodology is the trade-off between multiple factors influencing the final quality.
Eventually, the extracted passages are ranked by their total weights. The top-ranked passages, whose overall words number does not exceed the required size, are regarded as summary.

3. Compliance with the Genre "Summary"

We consider some types of passages as inadequate for including into summaries. In our opinion, the reasons for this can be:
- excessive length of a passage;
- passages represented by interrogative or exclamatory sentences;
- quoted passages;
- sentences in which the subject is expressed by a first-person pronoun.

Some other types of passages can be accepted as candidates for including into summary, though their features decrease suitability for the summary genre. For example, our premise is that indirect speech is less suitable in a summary than direct factual statements. Thus, passages with lexical-syntactic constructions introducing quoted and not quoted indirect speech are fined with negative scores.

4. Relevance

4.1. Terms identified while analyzing the topic statements

As indicators of relevance, we consider the words remaining in title/narrative text after excluding:
- syntactic and some other words from the stop-list;
- introductory parts of sentences such as “Describe the issues”, “Identify”, “Compare”, “Name”, etc. (we created the list of such expressions by analyzing topic descriptions in past-years DUCs and have added other possible variants);
- words/expressions that are interpreted as indicators of a focus requested for the respective summary. We distinguish these expressions from indicators of relevance and list them in special embedded vocabularies (see

Section 5).

4.2. Derivatives and co-occurring words

The derivatives of title/narrative words are also indicative of relevance. In our system, derivatives are created by adding or replacing suffixes. We also use as synonyms the different names of the same country and its derivatives (e.g., Britain, Britain's, British, UK, England, English).

The third group of terms indicative of relevance are non-stop-words most frequently occurring in the collection of documents covering the requested topic. We assume that these words are frequently co-occurring with title/ narrative terms and are semantically related to the topic.

Any term of these three groups, if present in a passage, is considered as influencing the level of its relevance. The extracted passages - candidates for including into the summary - must contain at least one of such terms.

4.3. Relevance scores assigned to different groups of terms

We have found that the relative role of "title" and "narrative" words as factors of relevance can be very different depending on their frequency in the processed documents. Because of this, the relevance scores assigned to title and narrative words are calculated so that increased scores are assigned to title words if their occurrence frequency is significantly less than overall frequency of narrative words - and vice versa.

Constant relevance scores are assigned to derivatives and co-occurring words.

All scores are multiplied by occurrence frequency of a given word. Additional scores are assigned to a passage for presence of multi-word phrases from the title. A passage also gets premium scores for presence of terms indicative of relevance from two and more groups outlined above (title words, narrative words, derivatives, co-occurring words).

4.4. Negative relevance scores

The system checks text passages for presence of phrases that share words with phrases in the topic statement but are semantically incompatible with them. The relevance scores of such passages can be decreased to zero.

Some examples of semantically incompatible phrases with shared words:
world war – civil war - Balkan war - war in Iraq;
new economy – economy of [a country] - world economy;
interest rates – growth rates – birth rates – mortality rates.

In the same way, as incompatible, we treat certain partial homographs:

"World" (e.g., as part of organization's name) – "world" (in any word combination).

4.5. The total relevance score of a passage

The shortness of a passage can be considered as a specific factor of relevance: the higher is overall number of words in the passage - the less is its relative relevance (determined by presence of the relevance-influencing words).

The sum of positive and negative relevance scores, correlated with the words number, characterizes total relevance of the passage.

5. Focusing

The topic statement can contain words/phrases characterizing the required focus of the summary that we interpret as pointing to logical categories by which the topic must be treated, e.g., "Causes", "Consequences", "Purpose", "Methods", "Advantages and Disadvantages", etc. For example, focus "Causes" can be expressed by use of such words/phrases as "cause" ,"reason", "impact", "factor", "accounts for", "leads to", "because of", and many others. The focus "Methods" can be expressed by "means of", "remedy", "techniques", 'via", etc.

We have 18 embedded vocabularies of focus terms. Most of the built-in focus lists used in our DUC system have been incorporated in our earlier products.

Passages with focus words/phrases get considerable positive scores. These scores are different depending on the following conditions:

- focus term and relevance term are adjacent to each other in the passage text (the highest score);
- focus term occurring in the passage is the same that is present in title/narrative (medium score);
- other focus term (the lowest score).

6. Granularity

We consider the plural forms of countable nouns from the title text and the presence of such words/expressions as "generally", "usually", "as a rule", etc. as indicative of "General" granularity.
As indicators of "Specific" granularity, we consider, in particular, mentioning of places and dates (including non exact dates such as "this Monday", "last month", "yesterday", "on Tuesday"). Passages are checked for the presence of such expressions before the latter are removed from the text (see Section 1).

To comply with the "General" granularity, positive scores are assigned for presence of its indicators and negative scores for indicators of "Specific". To comply with "Specific", scores assignment is mirroring to the above.

7. Ordering Passages Regarding the Criteria of Cohesiveness

Ordering of extracted sentences/passages was recognized as a challenging task first in the field of multi-document summarization. As Barzilay et al. stated: “The problem of organizing information for multi-document summarization so that the generated summary is coherent has received relatively little attention. While sentence ordering for single document summarization can be determined from the ordering of sentences in the input article, this is not the case for multi-document summarization where summary sentences may be drawn from different input articles” [3].

Our principal approach to discursive ordering of passages has been primarily developed to be used in building not short summaries, but extended free-style reports. This approach relies on theories of thematic progression and "given-new" (theme-rheme) relationships of text units [4, 5, 6]. Our system InformationCompiler performs the ordering at two levels. At the first level, passages are grouped into sections. Each succeeding section covers a more specific, less general topic than the preceding ones. At the second level, passages within the sections are ordered by lexical cohesion.

Our DUC-2005 system uses only modified cohesion criteria. Relevance and focus terms repeating in every following passage get positive scores that correlate with their scores in the preceding passage.

Thus, passages ranked by the sum of scores for relevance, conformity to the genre "summary", focusing and compliance with the requested granularity are rearranged at this step with respect to cohesion scores.

Passages with too many repeating words are considered as redundant and excluded from the summary.

8. Coverage

To ensure the increased coverage, we assign premium scores for each **new** appearance of title, narrative, and co-occurring words that were absent in the preceding passages.

There are also scores for diversity inside a specific class of entities. For example, if a topic must be covered "worldwide", each newly-mentioned country increases the positive score of a passage.

The coverage scores are taken into account at the stage of passages ordering.

At this stage, passages are ranked by the sum of their scores.

9. Balancing the Impacts of Different Text Units' Features

Efficiency of the described approach largely depends on relative significance of multiple lexical and grammatical features that are related to each factor of summary quality and involved in the scoring of candidate passages.

Achieving the trade-off between the relative impacts of each text feature was the most labour-consuming part of the system development.

Balancing the impacts of different text features in course of the initial "tuning" of the system was performed on empirical basis, using the sample text materials provided by NIST as well as documents collections gathered on the web. This work was aimed at detection of correlation between quality of the summary as a whole (assessed by human evaluation) and the quantitative data on relative contribution of every text feature to the text units scores.

10. The Obtained Results

The scores obtained by our system (FTextST-05, assigned ID = 5) according to different evaluations (responsiveness, Rouge, linguistic qualities) show that our system is positioned among 2-5 that turned to be most effective (20.79 average responsiveness).

However, we realize that the applied method, its multiple specific rules, parameters, and vocabularies require further testing and improvement.

First of all, we feel that multiple factors might be better balanced. A number of our summaries bear evidence of conflicting relationship between textual coherence and coverage. The sequence of text units cohesively adjacent to preceding ones usually provides an impression of well organized text, but on the other hand, this sometimes leads to the lower diversity of the content and to decreased coverage. And vice versa - covering too many aspects of the topic results sometimes in poorly organized text.

A shortcoming of some of our summaries is the failure to comply with requested granularity, because of insufficient set of means affecting this factor of quality.

11. Possible Implications to Summaries Evaluation

The summarization approach that relates features of text units to the final quality of the summary suggests new insights to the summary evaluation task.

In fact, this model incorporates evaluation into the process of building the summary. Final summary is the sequence of text units that got the highest total rating while taking into account all factors influencing the quality, each represented by numerous text features.

Of course, there can be different implementations of this approach - with different choice of factors of quality, scoring methodology and so on. But we believe that such approach in general could serve, after necessary experimentation, as an instrument for evaluation of automatically composed summaries (regardless of the method applied for their creation).

12. Future Work

We intend to adapt our summarization technology to practical use in real-life tasks. Regarding this point, we are in accord with authors of [7].

We see as promising such directions for the future work as:
- producing a logically structured text, with sections and paragraphs (the latter capability is important for extended reports in contrast to the short summaries);
- editing summaries aimed at removing their non-informative parts (without distorting the grammatical correctness);
- prioritizing the salience and pertinence of information included in summary (in addition to relevance);
- customizing the summarization model for specific domains.

To some extent, these directions are implemented in our existing products and online demos [1].

A noticeable step further in the direction of the real-life tasks would be coherent arrangement of topic-related information items gathered from unsorted documents (large corpora of newspapers articles, etc.) instead of prebuilt relevant collections. Though in such case, the output report hardly could be called a "summary", we consider such task to be worth attention of future Document Understanding Conferences.

References

1. http://www.powertextsolutions.com

2. Martin Soubbotin, Sergei Soubbotin. Exhaustive Mining of Information from Unstructured Documents. The 9th World Multi-Conference on Systemics, Cybernetics and Informatics. Orlando, 2005.

3. R. Barzilay, N. Elhadad, K. R. McKeown. Inferring strategies for sentence ordering in multidocument news summarization. Journal of Artificial Intelligence Research, v. 17, 2002, pp. 35-55.

4. Danes, F. (ed.) 1974. Papers on Functional Sentence Perspective. The Hague: Mouton.

5. Fries, Peter H. 1983. On the Status of Theme in English: Arguments from Discourse. In: Petöfi, J. S. and E. Sozer. (eds.). Micro and Macro Connexity of Texts. Hamburg: Buske: pp. 116-152.

6. Petöfi, Janos (ed.). 1988. Text and Discourse Constitution. Berlin: Gruyter.

7. E. Amigo, J. Gonzalo, V. Peinado, A. Penas, F. Verdejo. An Empirical Study of Information Synthesis Tasks, nlp.uned.es/pergamus/pubs/articuloACL2004.pdf